Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 15, Number 4, May, 2010 ISSN 1531-7714

Improving Grading Consistency through Grade Lift Reporting

Ido Millet
Pennsylvania State University, Erie

We define Grade Lift as the difference between average class grade and average cumulative class GPA. This metric provides an assessment of how lenient the grading was for a given course. In 2006, we started providing faculty members individualized Grade Lift reports reflecting their position relative to an anonymously plotted school-wide distribution. Two schools elected to participate in this reporting, and two other schools declined. To analyze the effect of Grade Lift reporting, we used paired comparisons of Grade Lift measures for the same faculty teaching the same course before and after reporting has started. Statistical analysis shows that, only in the two schools that participated, there was a reduction in both variance as well as average levels of Grade Lift. If these results can be replicated at other universities, Grade Lift reporting may become a useful tool for increasing grading consistency.

While grade inflation has been the topic of much discussion (e.g. Goldman, 1985; Cole, 1993; Gradeinflation.com, 2009; Johnson, 2003; Schiming, 2009), this paper aims at improving grading consistency across faculty members. One reason that not much has been done about grade inflation (Rojstaczer, 2009) is that a unilateral lowering of grades might hurt the prospects of our own students. In contrast, lowering grading variability across faculty members may actually benefit our students by facilitating unbiased choices of electives and areas of study (Felton & Koper, 2005).

One approach for lowering grading variability across faculty member is to use common examinations (Bond, 2009). However, as acknowledged by Bond, "programs that regularly employ common examinations are still rare, primarily because they require a significant investment of faculty time and effort." Another approach for increasing grading consistency is to create formal guidelines for the distribution of grades. For example, Princeton University's grading guidelines "posit a common grading standard for every academic department and program, under which A's (A+, A, A-) shall account for less than 35 percent of the grades given in undergraduate courses" (Malkiel, 2009). Again, perhaps due to the focus on curbing grade inflation, such approaches are rare. Furthermore, while they may increase grading uniformity across departments, they do not address grading variations across faculty within departments.

Focusing our attention on improving grading consistency across faculty members, we may accept the existing level of grades as a norm for our school, but aim to reduce faculty variations around that norm. This, of course, begs the question of how these grading variations should be measured and communicated.

Since differences in student performance are a legitimate source of variations in grades, we need a way to separate and measure the effects of instructor grading leniency. This paper describes a Grade Lift reporting system as a way to measure and highlight faculty deviations from grading norms, after factoring out a proxy measure for student performance.

For the last three years, we have used this reporting system at two of our four schools. Having two participating schools and two non-participating schools created a perfect benchmarking situation whereby the effects of the reporting system on the two participating schools can be compared to the two non-participating schools. To further control for faculty and course effects, we used paired comparisons contrasting grading profiles of the same faculty member, teaching the same

course, one year before the introduction of the reporting system and two years later. This analysis shows that only the participating schools achieved an improvement in grading consistency.

The paper starts by describing the Grade Lift metric and the reporting system. This is followed by results from an anonymous survey of faculty reactions to the system. We then provide statistical analysis using paired comparisons of Grade Lift before and after reporting had started for the two sets of schools. We conclude with implications for practice and future research.

RELATED LITERATURE

One important concept in this reporting system is the Grade Lift metric. We will review the technical and organizational aspects of the system after first describing that idea.

Grade Lift Metric

The Grade Lift metric measures the difference between the average grade a class received and the average Cumulative GPA (CGPA) of the class students at the end of that semester. The CGPA is taken at the end of the semester since it best reflects the student quality at that point in time. This also ensures that even first-semester and transfer students have a reference CGPA.

The intuitive appeal of the Grade Lift metric comes from its ability to remove student quality from the discussion and focus attention on faculty grading behavior. Consider, for example, an instructor who assigned an average grade of 3.0 to her class. If the average CGPA of the students in this class was 3.5 (Grade Lift = -0.5), we may conclude that this instructor graded students in that class more harshly than institutional norms. Conversely, if the average CGPA of the students in this class was 2.5 (Grade Lift = +0.5), we may conclude that this faculty member grades more leniently than the norm.

Superficially, the core idea for the metric may seem similar to prior literature advocating the use Class GPA as a useful reference for student grades. For example, Felton & Koper (2005) suggest using class GPA in order to compute *real* (as opposed to *nominal*) student GPA. The difference is that prior approaches focus on better measurement of *student quality* by benchmarking student grades against class GPA. In contrast, the Grade Lift metric focuses on measurement of *instructor grading behavior* by benchmarking instructor grades against class CGPA. In short, the Grade Lift metric measures grading leniency by subtracting expected grades (based on student CGPA) from assigned grades.

Grade Lift Metric Limitations and Sources of Grading Variations

Obviously, the Grade Lift metric is not a perfect measure of grading leniency. A major limitation of this metric is that average CGPA for a group of students in a given class is only a very rough proxy for their average performance in that class. Instead, it's a measure of their *past* performance on *different* courses. Furthermore, there are many sources of Grade Lift variations across faculty members, and some of them are quite legitimate.

One example of a legitimate source of variation is the instructor's ability to motivate and inspire good work by the students. Another legitimate source of variation is changes in students' levels of interest, motivation, and even ability as they progress through their program of study. For example, engineering students may struggle with the quantitative first-year courses and continue to perform poorly in similar second-year courses (low or negative Grade Lift) but excel (high Grade Lift) in more applied second-year courses. In other words, instructors who teach groups of students whose past courses require different aptitudes and interests would probably experience higher variations in Grade Lift.

Still, these same effects would also frustrate attempts to use simple grade average metrics as a measure of faculty grading leniency. On the other hand, several illegitimate sources of grading variations are captured by the Grade Lift metric. We know from prior research (Sonner, 2000) that lower ranking faculty are more lenient graders and that some disciplines, such as Economics, are tougher graders than others.

It should be noted that some sources of grading variations are difficult to classify with respect to legitimacy. For example, smaller class sizes have been shown to be correlated with higher grades (Sonner, 2000). On the one hand, that effect could be due to legitimate reasons such as higher student interest in the topic and a better learning environment. On the other hand, as Sonner (2000) explains, this could be due to a leniency bias introduced by closer working relationships.

To address the limitation of interpreting results from very small classes, and to ensure Grade Lift metrics are based on enough observations, our reporting system includes only course sections with more than 5 students.

This also provides a partial solution for the biasing effect that failed grades could introduce.

Failed grades, as one anonymous reviewer observed, can introduce a strong biasing effect on average grades. We considered removing failed grades from the computation of Grade Lift metrics but our administration elected to keep them in. The reasoning was that this keeps thing simple and aligned with the way GPA metrics are computed. Obviously, other schools may elect to remove failed students from the metric.

Another limitation of the Grade Lift metric is due to the timing of measurement. First-semester, part-time, or transfer students may have a CGPA based on very few courses. In the most extreme case, a student who has taken his first and only course with us would end up with a CGPA that is identical to his assigned grade. This means that we should treat the Grade Lift metric with more caution when applied to instructors who teach lower-level courses. It also means that this metric may not be a good tool for short degree programs.

While we have been using the Grade Lift metric with good results, other universities could use a simple average grade metric with the same reporting approach described in the following section. *The key idea is that when instructors know how their grading profiles compare to their peers, extreme grading behaviors begin to moderate.*

Report Design and Distribution

Our system draws student, course, instructor, and grades data from an institutional data warehouse. The same data should be easy to obtain in most academic institutions. At the end of each semester, an automated process emails individualized reports, so that each faculty member gets only their own information.

Figure 1 depicts one key display in the report. The bars correspond to the Grade Lift for each anonymous instructor, averaged across their course sections in the last semester. We elected to use a simple average, ignoring section size. A weighted average (by section size) could easily be used as an alternative.

As shown in Figure 1, the vertical bar corresponding to the instructor receiving the report is highlighted. This makes it easy for each instructor to see how they compare to the anonymous Grade Lift distribution of their peers. In this particular case, the faculty member ranks fifth out of 45 faculty members on the Grade Lift metric.

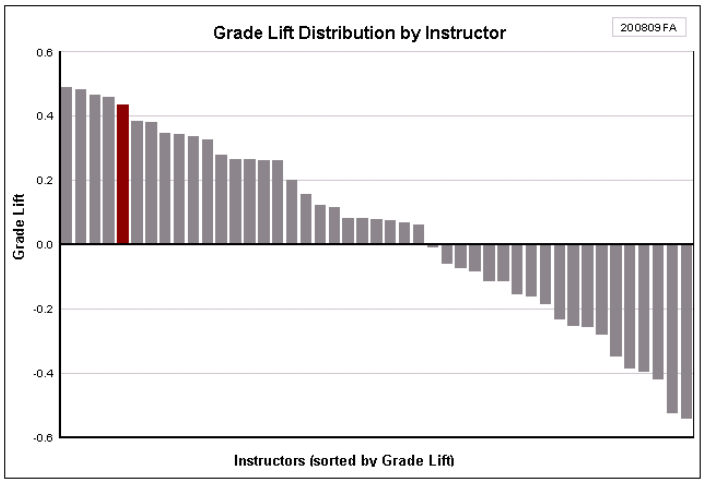

Figure 1. Anonymous Chart of Average Grade Lift by Instructor

The shape of the distribution conveys useful information. For example, more instructors are on the positive side but there are many on the negative side as well. The slope of the distribution is relatively constant, indicating that variability is not limited to a few extreme cases. We can see that there is a full grade point difference between the highest (+0.49) and lowest (-0.54) average grade lift per faculty member. To put this within our 4.0 grade scale perspective, the instructor with a +0.49 average Grade Lift tends to give B+ students A or A- grades, while the instructor with a -0.54 average Grade Lift tends to give B+ students only B- grades.

The report also shows for each faculty member the distribution of grades and Grade Lift in each section, and how the Grade Lift compares to average Grade Lift for courses at that level.

For example, in the case shown in Figure 2, the first course section is a 300-level (3rd-year) course, and the Grade Lift of 0.55 is significantly higher than the -0.02 average for all 300-level courses in the school that semester.

These Grade Lift benchmarks and faculty distributions are isolated for each school. This ensures that faculty members are compared only to their close peers.

The reports also include summary charts and cross-tabulations for several semesters. This allowed us to spot interesting patterns. For example, Grade Lift tends to be higher during summer semesters. One possible explanation is that summer semesters are typically staffed by a higher proportion of adjunct teachers who tend to give higher grades (Sonner, 2000).

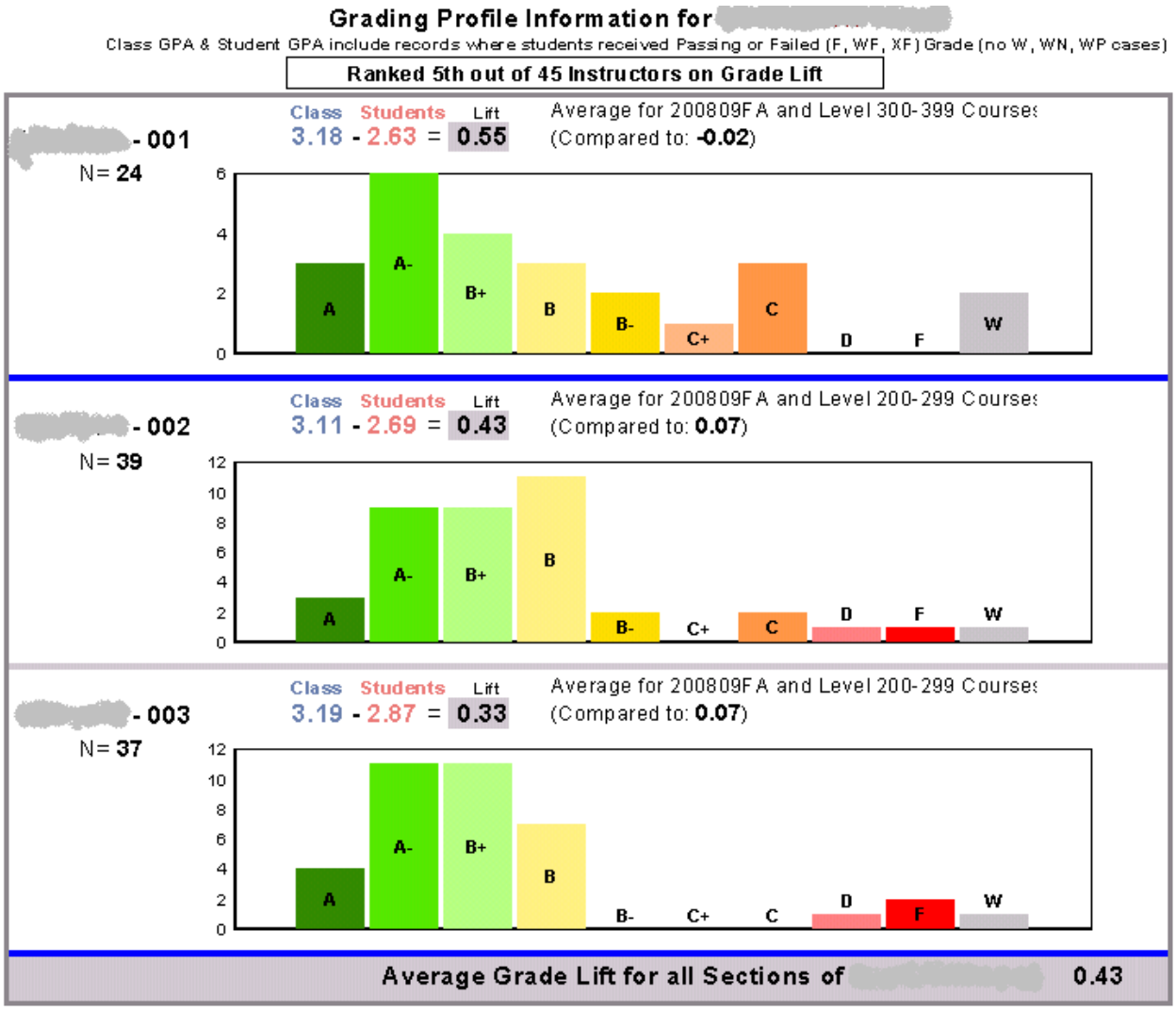

Figure 2. Grade Distribution & Lift Benchmarks by Course Level

Organizational Approach

Our school directors avoided wielding Grade Lift information as a way to single out or browbeat faculty members. Instead, they elected to introduce the information as feedback to faculty in an anonymous and non-threatening way. At the end of each semester, we electronically burst and email the reports so that each faculty member gets a clear picture of their own grading profiles. One school director includes the following in the automated email message:

> The information in the attached pdf file shows the grade distribution and Grade Lift for each of your course sections. Besides the grade distribution chart, you can see how your Grade Lift compares to the school's average for courses at that level. You can also see how your average Grade Lift (across all your sections) ranks among other faculty members. The last 2 pages of the report show overall trend information on Students GPA, Class GPA, and Grade Lift by course level for the school.
>
> I hope that this report prompts at least a few moments of review of your own grade distributions as well as conversation between you and your colleagues about expected levels of achievement associated with various letter grades. Over time, I hope such thought and discussion lead to greater consistency in grading across sections and faculty.

The system was first implemented in 2006 and each of the two participating schools has about 45 faculty members.

Faculty Reactions

We used anonymous questionnaire in both participating schools to collect faculty opinions about the system. We received 23 responses (26% response rate). While the real evidence for the system effectiveness is described in the next section, faculty responses provide a useful indication for how well the system was received.

In response to the question of "should we continue Grade Lift reporting" 18 faculty members (78%) responded with a Yes, two abstained, and three responded with a No. Most instructors seem to welcome the system. Examples of *positive comments* include:

> "I think the feedback is helpful. I noticed lift this past semester and I will take steps now to contain it."
>
> "I like it. Please continue using it. I think it provides valuable feedback for the instructor."
>
> "I find it very helpful!"
>
> "We definitely need to do something like this to help combat the natural tendency to grade inflation that has infected post-secondary education"
>
> "It is a useful bit of feedback and is an easy way for faculty to sense whether their grading standards are reasonable."

Examples of *negative comments* include:

> "For this system to work there has to be some accountability when people are shown to have grade lift. Currently, there is no reward or punishment and there is an obvious correlation between grade lift and teaching evaluation scores. So why should anyone voluntarily become a tougher grader?"
>
> "Unless there are consequences, these reports will do nothing."

Only 13 (57%) out of 23 responding faculty members indicated the system will actually nudge faculty members closer to school norms. We can conclude that while most faculty members see value in the reports and want them continued, there is a significant minority that is skeptical about actual effects. That significant minority (43% of respondents) may be pleasantly surprised by the positive evidence provided in the following section.

RESULTS

In order to isolate the effect of the system on Grade Lift measures, we need to remove noise such as changes in faculty and course assignments. We achieved this by selecting only cases where the same faculty member taught the same course and same section in the year prior to starting the reports, and two years later. We had 186 such cases.

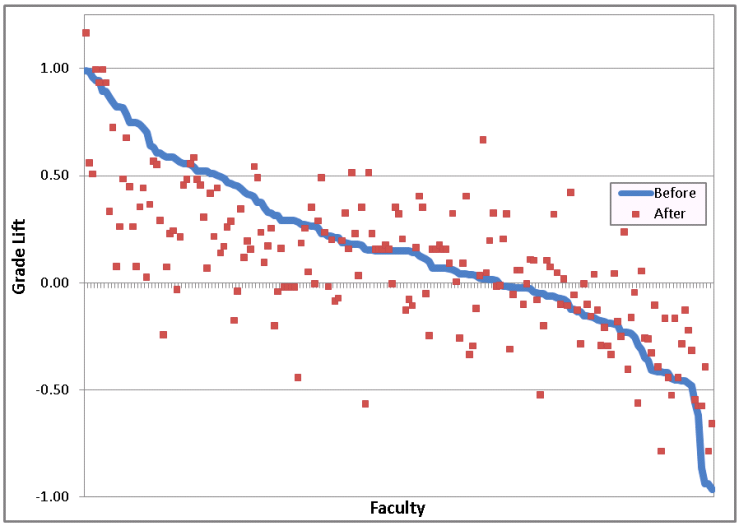

Figure 3: Instructors with Extreme Grade Lifts Moved to the Center

Figure 3 shows that, in the two participating schools, Grade Lift has moderated, particularly among those faculty members who started with extreme positive or negative measures. Since the chart is sorted by initial Grade Lift (the blue line), instructors with high initial grade lift are at the top left corner of the chart. We can see that the majority of these instructors (26 out of the top 30) moved to lower Grade Lift measures two years later. Similarly, the majority of the toughest graders (24 of the bottom 30) at the bottom-right corner of the chart moved to less negative Grade Lift measures. While these changes among the extreme cases could well be due to the effect of regression to the mean, the chart also shows that most intermediate cases did not evolve new extreme behaviors. In other words, not only the extreme cases, but the overall distribution seems to have moderated. This visual impression is confirmed by the following statistical analysis.

Table 1 shows results of an F-test indicating that in the two participating schools there was a statistically

significant (α = 0.009) reduction in Grade Lift variance from 0.151 to 0.107.

Table 1: F-Test for Variance (participating schools)

	Before	*After*
Mean	0.146687	0.081149
Variance	**0.151406**	**0.106934**
Observations	186	186
Df	185	185
F	1.415891	
P(F<=f) one-tail	**0.009238**	
F Critical one-tail	1.274414	

Table 2 shows results of a Paired t-Test confirming that in the two participating schools there was a statistically significant (α = 0.001) reduction in average Grade Lift from 0.147 to 0.081.

Table 2: Paired t-Test (participating schools)

	Before	*After*
Mean	**0.146687**	**0.081149**
Variance	0.151406	0.106934
Observations	186	186
Pearson Correlation	0.736228	
df	185	
t Stat	3.354894	
P(T<=t) two-tail	**0.000963**	
t Critical two-tail	1.97287	

Results for the Two Non-Participating Schools

Repeating the same analysis for the two non-participating schools shows that, during the same period, no significant changes in grade lift variability occurred in those schools. Table 3 shows results of an F-test indicating that **in the non-participating schools Grade Lift variance did not decrease**. In fact, the variance in these schools actually increased from 0.15 to 0.16, but that increase was not statistically significant (α = 0.31).

Table 3: F-Test for Variance (non-participating schools)

	Lift_Before	*Lift_After*
Mean	0.134601	0.168125
Variance	**0.154807**	**0.163969**
Observations	308	308
Df	307	307
F	0.944123	
P(F<=f) one-tail	**0.307382**	
F Critical one-tail	0.828578	

Table 4 shows results of a Paired t-Test indicating that in the non-participating schools average grade lift actually increased from 0.13 to 0.17. That increase was statistically significant (α = 0.03) and in the opposite direction to what happened in the two participating schools.

Table 4: Paired t-Test (non-participating schools)

	Lift_Before	*Lift_After*
Mean	**0.134601**	**0.168125**
Variance	0.154807	0.163969
Observations	308	308
Pearson Correlation	0.769796	
df	307	
t Stat	-2.17034	
P(T<=t) two-tail	**0.030748**	
t Critical two-tail	1.967721	

DISCUSSION

We may conclude that Grade Lift variance and level have been reduced, probably due to the introduction of the Grade Lift reporting system, in the two participating schools. The chart in Figure 3 and the Paired t-Test Pearson Correlation of 0.74 (Table 2) indicate that instructors retained their relative grading tendencies. Most lenient graders remained lenient, and most tough graders remained tough. The system seems to moderate very lenient or very tough grading by providing feedback and awareness of norms.

Some of the faculty comments described earlier indicate a general belief that lenient grading results in better student evaluations. Prior studies of such a relationship (Franklin, 1991) were limited by the fact that student quality was not factored out from the grade metrics. The Grade Lift metric provides an opportunity

to investigate the relationship between grading leniency and student evaluations. Furthermore, the same paired-comparison approach used in this paper may prove valuable in removing noise due to faculty and course changes.

It should be noted that two courses with the same grade lift may have vastly different internal dynamics. One course with zero Grade Lift may have individual grades that perfectly match each student's CGPA. Another course with zero Grade Lift may have assigned low grades to students with high CGPA, and high grades to students with low CGPA. We can develop new "Grade Alignment" metrics based on the distance between individual grades and CGPA. Such metrics may provide useful diagnostics alerting faculty members to review the assessment methods used in certain courses. We can expect Grade Alignment metrics to have positive correlations with student evaluations of teaching since a good student's resentment of a poor grade may be stronger than a poor student's welcome of a good grade.

As one anonymous reviewer noted, pursuing extreme grading consistency would lead to obviously dysfunctional goals. An issue that remains to be resolved is what level of grading inconsistency is acceptable. Collecting grade lift information from multiple schools may provide interesting benchmarks.

Given the positive results we have witnessed, we are continuing to use the system at our two participating schools. We hope this paper helps other schools implement a similar Grade Lift reporting system as a way to inform faculty members and improve grading consistency.

REFERENCES

Bond, L. (2009). The Case for Common Examination. The Carnegie Foundation for the Advancement of Teaching, Retrieved October 17, 2009 from: http://www.carnegiefoundation.org/perspectives/sub.asp?key=245&subkey=2207

Cole, W. (1993). By Rewarding Mediocrity, We Discourage Excellence. *The Chronicle of Higher Education,* XXXIX(18), (January 6, 1993), B3-B4.

Felton J. & Peter T. Koper T. P. (2005). Nominal GPA and Real GPA: A Simple Adjustment that Compensates for Grade Inflation. *Assessment & Evaluation in Higher Education*, 30(6), 561-569.

Franklin J., Theall M., & Ludlow L. (1991). Grade Inflation and Student Rating: A Closer Look, paper presented at the American Educational Research Association, Chicago, IL, April 3-7. Retrieved October 17, 2009 from: www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/38/2b/dd.pdf

Goldman, L. (1985). The betrayal of the gatekeepers: Grade inflation. *Journal of General Education*, 37, 97-121.

Gradeinflation.Com. (2009). *Grade Inflation at American Colleges and Universities*. Retrieved October 17, 2009 from: www.gradeinflation.com

Johnson, V. (2003). *Grade inflation: A crisis in college education*. New York: Springer-Verlag.

Malkiel, N. (2009) Grading Policy Letter, *The Daily Princetonian*, Retrieved October 17, 2009 from: http://www.dailyprincetonian.com/2009/04/23/23516/

Rojstaczer, S. (2009), Grade inflation gone wild, *Christian Science Monitor*, March 24, 2009, Retrieved October 17, 2009 from: www.csmonitor.com/2009/0324/p09s02-coop.html

Schiming, R. (2009), Grade Inflation Article, Retrieved October 17, 2009, from: www.mnsu.edu/cetl/teachingresources/articles/gradeinflation.html

Sonner, B. (2000). A Is for "Adjunct": Examining Grade Inflation in Higher Education. *Journal of Education for Business,* 76(1), 5-8. Retrieved October 17, 2009, from: http://jan.ucc.nau.edu/~slm/AdjCI/Evaluation/Inflation.html

Citation

Millet, Ido (2010). Improving Grading Consistency through Grade Lift Reporting. *Practical Assessment, Research & Evaluation*, 15(4). Available online: http://pareonline.net/getvn.asp?v=15&n=4.

Corresponding Author:

Ido Millet
Professor of MIS
Penn State Erie, Black School of Business
Erie, PA 16563-1400
email: ixm7 [at] psu.edu